Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects

Stefano Breschi ¹, Francesco Lissoni ¹/²/§

¹ CRIOS – Università Bocconi, Milan
² GREThA – Université Montesquieu, Bordeaux IV
§ contact author: francesco.lissoni@u-bordeaux.fr

This version: 10 June 2014
To be presented at: 9th EPIP workshop “Improving Innovation Systems” – September 4-5, 2014, Brussels – European Parliament

Abstract

We assess the role of ethnic ties in the diffusion of technical knowledge by means of a large database of patents filed by US-resident inventors of foreign origin (“ethnic” inventors), 1980-2010. Ethnic inventors are identified by means of linguistic analysis of their names and surnames, with reference to ten important countries of origin of highly skilled migration to the US (China, India, Iran; Japan, S.Korea for Asia; France, Germany, Italy, Poland, and Russia for Europe). We test whether ethnic inventors’ patents are disproportionately cited by: other migrants in the same destination country (the US) and from the same country of origin (“diaspora” effect); and non-migrant inventors residing in their country of origin (“brain gain” effect). We find strong evidence of both the diaspora and the brain gain effect for China and India, and some weaker evidence for almost all other countries in the sample. Cross-country differences, however, may be in part explained by data quality issues, which affect especially inventors of European origin. For foreign inventors in the US, physical and social proximity appear to be more important determinants of citation flows than co-ethnicity. The same applies, when considering international knowledge flows, to organizational proximity (affiliation to the same multinational company) and, again, to social proximity. It remains to investigate the role of ethnic ties in the formation of inventors’ networks.

Keywords: migration, brain gain, diaspora, diffusion, inventors, patents

JEL codes: F22, O15, O31

ACKNOWLEDGEMENTS: Unique identifiers for inventors in the EP-INV database come from the APE-INV project (Academic Patenting in Europe), funded by the European Science Foundation. The pilot project for assigning inventors to ethnic group was funded by the World Intellectual Property Organization (WIPO), which also made available to us the WIPO-PCT dataset (with Ernest Miguélez providing key help). Discussants at the first WIPO Experts Meeting on IP and the Brain Drain (Geneva, April 2013) provided useful suggestions. The same apply to participants to the following conferences: MEIDE (Santiago de Chile, November 2013), PATSTAT (Río de Janeiro, November 2013), EUROLOIO (Utrecht, January 2014); as well to participants to seminars at University College Dublin, LSE, CRIOS-Bocconi, Kassel University, Collegio Carlo Alberto in Turin, and IMT in Lucca. Gianluca Tarasconi provided outstanding technical assistance and silly jokes. We owe the discovery of the IBM-GNR© system to Lars Bo Jeppesen, while Curt Baginski assisted us in its implementation.
1. Introduction

The last decade has seen the convergence of two important streams of literature dealing with the diffusion of technical knowledge and the mobility of scientists and engineers. First, research in the geography of innovation has explored the role of social ties in facilitating knowledge diffusion, and in determining its spatial reach. Among such ties, a good deal of attention has been paid to those binding members of scientific and technical “diasporas”, namely the communities of migrant scientists and engineers from the same origin country (Agrawal et al., 2008; Kerr and Lincoln, 2010; Saxenian et al., 2002). Second, migration and development scholars have explored to what extent these diasporas contribute to innovation in their home countries, through international knowledge flows (Kapur, 2001; Kuznetsov, 2006; Saxenian et al., 2002). Emerging naming conventions label the social ties in question as “ethnic”, a synthetic but imperfect adjective we will also adopt, for want of better alternatives.

The two streams of literature share a common necessity in going beyond anecdotal evidence and success stories. This requires measuring in general terms the actual importance of ethnic ties as vehicles for knowledge diffusion, and in assessing the relative weight of their multiple embodiments. The latter comprise multinational firms operating in both the destination and home countries of migrants (Foley and Kerr, 2011), several academic and professional exchange networks (Meyer, 2001; Meyer and Brown, 1999), as well as returnee migration, when returnees participate to the home country labour market (Alnuaimi et al., 2012; Nanda and Khanna, 2010) or to entrepreneurial ventures (Saxenian, 2006; for a skeptical view, see: Kenney et al., 2013).

Patent and inventor data have been increasingly used to address these measurement issues. Migrant inventors are identified as such either by using information on their nationality, available on Patent Cooperation Treaty (PCT) applications (Miguelez, 2014; Wadhwa et al., 2007b); or by means of a linguistic analysis of their names and surnames (Agrawal et al., 2008; Agrawal et al., 2011; Almeida et al., 2010; Foley and Kerr, 2011; Kerr, 2008; Kerr and Lincoln, 2010). So far, however, both streams of literature have focussed almost exclusively on the US as a destination country, and on China and India as origin countries of highly skilled migrants in general, and migrant inventors in particular. This overlooks the fact that several European countries are also important sources of highly skilled migrants to the US; and that Europe hosts quite robust flows of intra-continental migration (Docquier and Marfouk, 2006; Widmaier and Dumont, 2011). The focus on China and India is also at the origin of present difficulties in assessing whether the evidence on the role of those countries’ diasporas can be generalized to other countries (Pandey et al., 2006).

In this paper we contribute both substantively and methodologically to this emerging field by analysing the forward citation patterns of patents filed by foreign inventors in the US from five Asian countries and as many European ones. All our data are novel and come from EP-INV, a database of uniquely identified inventors listed on patent applications filed at the European Patent Office, combined with information from IBM-GNR© (Global Name Recognition system, courtesy of IBM), by means of an original algorithm. Complementary (benchmarking) data come from PCT applications, as made available by Miguelez and Fink (2013).

We use these data to test for the existence of “diaspora” and “brain gain” effects. We state a diaspora effect to exist when foreign inventors of the same ethnic group and active in the same country of destination (in our case, the US) have a higher propensity to cite one another’s patents, as opposed to patents by other inventors, other

---

1 Patent data have also become increasingly attractive for a substantial reason, namely the adoption of new, US- and Europe-like patent legislation by several inventors’ home countries, following their entry into the WTO and subscription of TRIPs. Research has just started on whether such strong patent legislation may favour FDIs and returnee entrepreneurship, thus compensating for making imitation of advanced countries’ technology more costly and we do not deal with it in this paper. (For a first attempt to organize a research programme on it, see the expert workshop organized by the World Intellectual Property Organization on “Intellectual Property, the International Mobility of Knowledge Workers, and the Brain Drain (April 29-30, 2013; http://www.wipo.int/edocs/mdocs/mdocs/en/cdip_12/cdip_12_+∞_5.docx)
things being equal. We state a “brain gain” effect to exist when the same foreign inventors’ patents are also
disproportionately cited by inventors in the foreign inventors’ countries of origin. We find evidence of both effects,
but not for all the countries investigated. We also find evidence that neither effect is quantitatively more important
than spatial and social proximity, the former measured by co-location at the metropolitan or county level, the latter
by geodesic distances on network of inventors. We also find that multinationals play an important role in mediating
international knowledge flows. Several results, however, are sensitive to the quality of inventor names’
disambiguation and company names’ harmonization.

In what follows, we first survey the literature on migration, innovation, and knowledge flows, with special
emphasis on patent-based studies (section 2). We then present our research questions and data (section 3). In
section 4 we report the results of our empirical exercise. Section 5 discusses such results and concludes.

2. Background literature

2.1 Localized knowledge flows and the role of social ties

Localized knowledge flows are a key topic in the geography of innovation (surveys by Breschi and Lissoni, 2001;
Breschi, 2011). Under the form of pure externalities (localized knowledge spillovers, or LKS), they play a key role in
Marshallian and Jacobian location theories (Ellison et al., 2007; Henderson, 1997). Yet, their importance has been
questioned both by New Economic Geography models (Krugman, 1991 and 2011) and by evolutionary theories of
clustering (Boschma and Frenken, 2011). A key point of contention in the debate has been that of measurement,
which is fraught with technical as well as conceptual difficulties.

As for technical difficulties, these were first tackled by Jaffe et al. (1993), who introduced the use of patent
 citations along with a simple, yet attractive methodology for testing their localization in space (from now on, JTH
test). The test makes use of two set of patent pairs. The first set includes a sample of cited patents and all the
related citing ones, with exclusion of self-citations at the company level (cited-citing or “case” pairs); the second
set includes the same sample of cited patents, with citing ones replaced by controls with the same technological
classification and priority year (cited-control or “control” pairs). After geo-localising patents on the basis of their
inventors’ addresses, a simple test of proportions is carried out, one that proves the share of co-localized cases to
be significantly higher than the share of co-localized controls (with co-localization specified either at the city, state,
or country level). The test can be generalized to a logit/probit regression, with the probability of a citation to occur
as the dependent variable, and the stacked sets of cited-citing and cited-control patent pairs as observations
(Singh and Marx, 2013).

Time and again, the JTH test has been successfully replicated, but also proven to be sensitive to the level of detail
chosen for the technological classification of patents (Henderson et al., 2005; Thompson and Fox-Kean, 2005), and
to the origin of patent citations (whether inserted in the patent documentation by the patent examiners or the
applicants; Alcacer and Gittelman, 2006; Thompson, 2006; see also Breschi and Lissoni, 2005, on citations attached
to EPO patents).2

On the conceptual side, the JTH test says nothing on the actual mechanisms behind localized knowledge flows
and their economic characteristics. Breschi and Lissoni (2005, 2009) show that a large share of localized patent
citations are self-citations at the individual level, associated to inventors who move (locally) across firms, or act as
consultants for different firms in the same location or region (as it is the case with academics or free-lance

---

2 EPO=European Patent Office. All studies mentioned in this section made use of USPTO (US Patent & Trademark Office) data, with the only
exception of Breschi and Lissoni (various years), which used EPO data.
Many others are citations between patents signed by socially close inventors, namely inventors located at short geodesic distances on networks of inventors. These results have two implications:

1. They cast some doubts on the pure externality (spillover) interpretation of localized knowledge flows: market mechanisms may be in place, such as labour mobility and/or licensing. It may also be the case that mergers and acquisitions or corporate restructuring are at work, with some firms’ knowledge incorporated into new entities, along with their inventors (see section 2.3).

2. They suggest that spatial proximity is largely a proxy for social proximity, in this case declined as professional proximity (who worked with whom). Professional proximity may force inventors to share knowledge (as when work together on the same project) or create strong enough social bonds to induce obligations or opportunities to share (as when two inventors who do not work together exchange information, voluntarily or not, through a common colleague or a short chain of colleagues). Agrawal et al. (2006) show that professional ties may indeed resist to physical distance, as when new patents by inventors who relocated keep being cited by former co-inventors who did not.

This line of research has evolved in the direction of uncovering other forms of social ties besides the professional ones, and of exploring their relationship with spatial distance. This has led Agrawal et al. (2008) to explore the role of ethnic ties in the US-resident population of Indian inventors, often described in the literature as closely-knit “diaspora” (Kapur, 2001).

First, based on an Indian surname database, the authors identify ethnic Indian inventors of USPTO patents, all of them resident in the US (1981-2000). Second, they apply the JTH methodology and extend it to test not only the extent of knowledge flows’ co-location (at the MSA level), but also the importance of ethnic ties. Co-ethnicity of two patents’ inventors is found to be associated to an increase in the probability to observe a citation link between such patents. Besides, co-ethnicity and co-location seem to act as substitutes, with Indian inventors in the US activating their ethnic connections to reach knowledge assets located outside their metropolitan area. The sociological or organizational origins of ethnic ties (whether they derive from previous common study or work experiences and/or by affiliation to formal organizations of expat professionals or entrepreneurs) are not explored.

Almeida et al. (2010) also rely on an ad hoc collection of surnames to identify Indian inventors in the US (in the semiconductor industry). They find both evidence of intra-ethnic citations, as well as some indications that reliance on such citations is correlated to inventors’ productivity.

Agrawal et al. (2011) extend the Agrawal et al.’s (2008) data and methodology to the case of international knowledge flows and find that patents by Indian inventors in the US do not seem to attract a higher-than-average rate of citations from the inventors’ home country. The only (weak) exceptions are patents in Electronics, and patents owned multinational firms. Overall, these results go in the direction of suggesting that the Indian diaspora is not a major source of knowledge feedbacks for its country of origin. As suggested by Singh and Marx (2013),

---

3 Academic scientists who produce patentable inventions do not always disclose them to their universities’ technology transfer offices, either in the attempt to exploit them individually or as a result of contractual arrangements with industry sponsors. In both Europe, the US, and Japan, academic patents are a non-negligible share of total domestic patents, especially in science-based fields (survey by Lissoni, 2012). Academic inventors play a key role in bridging gaps and shortening distances in inventors’ networks (Lissoni et al., 2013).

4 Networks of inventors are obtained by examining co-inventorship patterns. Technically, they are one-mode projections of “two-mode” or “affiliation” networks (Borgatti and Everett, 1997). All co-inventors are at distance one, while inventors who have never worked together on the same patent, but have at least one co-inventor in common, are at distance two; if they have no common co-inventors, but at least two of their co-inventors once worked once together, they are at distance three; and so forth. When examining complex networks, in which any two inventors can be linked one another by several different “chains” of co-inventors (paths), the social distance between any two inventors is usually measured as the length of the shortest path between the two (geodesic distance).

5 One important difference between Agrawal et al.’s (2008) method and JTH is that the former makes use, as observations, of inventor pairs, rather than patent pairs. Whenever a cited-citing patent pairs include n>1 inventors in the cited patent and m>1 inventors on the citing one, this produces m*n observations. Another difference is that controls are selected on cited patents, rather than citing ones.
country boundaries seem to provide obstacles to knowledge diffusion that resist to controlling for spatial distance, that is are additional to the latter.

It is at this point that studies in the geography of innovation tradition blend with research on migration and brain gain.

2.2 Migrants’ contribution to innovation in origin countries

Migration studies have traditionally looked for possible positive returns from emigration for origin countries. Early research placed special emphasis on emigrants’ remittances and their role in capital formation. More recently, due to the increasing importance of highly skilled migration, more attention has been paid to emigrants’ contribution to knowledge stock building and innovation (Bhagwati and Hanson, 2009).

This may come in three, non-mutually exclusive forms, namely:

(i) “Ethnic-bound” knowledge flows. Emigrant scientists and engineers may retain social contacts with professional associations and educational institutions in their home countries, and transmit them the scientific and technical skills they have acquired abroad, either on a friendly or contractual basis (Meyer and Brown, 1999; Meyer, 2001)

(ii) Internal transfers by multinational companies.

(iii) Returnees’ direct contribution. Emigrant scientists and engineers who have worked as academic or industrial researchers, may decide to move back to their origin countries and continue their activities over there. In the case of entrepreneurs, they may keep base in the destination countries, but set up new or subsidiary companies in their home country (Wadhva, 2007a,b; Kenney et al., 2013, and references therein). 6

While case studies on these phenomena abound, large-scale quantitative evidence is scant and almost entirely focussed on the US as a destination country, and China and India as origin countries. This largely ignores the fact that highly skilled migration to the US originates not only from developing countries, but also from Western Europe, South Korea, and Japan (Docquier and Marfouk, 2006; Widmaier and Dumont, 2011; see also Freeman, 2010). 7

William Kerr, in a series of papers (some with co-authors), has exploited two sources of information:

- the NBER Patent Data File, by Hall et al. (2001), which includes information on name, surnames, and addresses of inventors

- the Melissa ethnic-name database, a commercial repository of names and surnames of US residents, classified by ethnicity, mainly used for direct-mail advertisements.

Names and surnames from the two sources are matched, so to assign each inventor to one out of nine broad ethnic groups, the most distinctive being the Russian, the Chinese (including Taiwanese and Singaporean), the Indian (including Pakistani and Bengali) and several other Asian ones.

As for knowledge flows, these are measured by citations running from patents filed at USPTO by inventors from outside the US, for each year between 1985 and 97, to patents filed by US residents over the 10 preceding years. Company self-citations are excluded, and the same applies to patents with less than 50% of the inventors either belonging to the same ethnic group or non-US country of residence. Citations are grouped according to four criteria (inventor’s ethnicity and technological class of the citing patent, plus inventor’s ethnicity and technological

---

6 In addition, the demand of highly skilled migrants by destination countries provides youth in origin ones with an incentive to get higher education. Such migration-induced demand for higher education allows origin countries to keep their university system going, even in fields for which local demand of graduates is lacking, and, with it, some absorptive capacity of foreign science and technology.

7 Besides, advanced European countries, while not as attractive as the US, host a non-negligible amount of highly skilled migrants from the East and South of the Old Continent, as well as from several former colonies.
class of the cited patent). A negative binomial regression is then run, with citation groups as observations, the number of citations in each group as the dependent variables (which is often zero), and a series of dummies as regressors. Among the latter, the “co-ethnicity” dummy is of particular interest, as it indicates whether the ethnicities of inventors in the cited and citing patents in the citation group are the same. This allows estimating that co-ethnic citation groups are on average 50% more numerous than mixed ethnic ones. This basic finding is interpreted as indicative of the existence of knowledge externalities (spillovers), very much along the lines followed by the LKS literature.

Kerr (2008) further uses patent data as regressors in a first-difference panel data econometric exercise concerning origin countries of immigration into the US, with economic growth as the dependent variable. Migrants’ patents in the US are found to affect substantially their home countries’ growth rates. The result weakens, but resists, when excluding China from the origin country set, or Computer and Drugs from the technologies considered. This suggests that ethnic-mediated spillovers, while having a stronger impact in high technologies and in one particular economy, are not irrelevant for a more general set of countries and technological fields. Notice that this positive results is in contrast with the weak evidence provided, for India, by Agrawal et al. (2011).8

Foley and Kerr (2011) exploit the same database to investigate the specific role of ethnic inventors in relation to multinational companies’ activities in origin countries. In particular, they find that US multinationals with a high share or quantity of ethnic patents invest and innovate more in their ethnic inventors’ origin countries, while at the same time relying less on joint ventures with local companies for doing so. This suggests that ethnic inventors may not only channel back to their origin countries some key economic and innovation activities, but also act of substitutes of local intermediaries, thus diminishing their companies’ costs of engaging into foreign direct investments.

As for returnee inventors, Agrawal et al. (2011) manage to identify very few of them, who are responsible of just 18 patents. Similarly, Alnuaimi et al. (2012) examine around 3500 USPTO patents assigned to over 500 India-located patentees (local firms, subsidiaries of foreign companies, and universities) in between 1985 and 2004, and find very few inventors once active in subsidiaries of foreign companies who then move to local firms. This suggests that, as far India is concerned, returnees and multinational employees in origin countries are not a direct source of knowledge transfer. Therefore, Foley’s and Kerr’s (2011) results can be only explained by indirect activities by ethnic inventors, not captured by patents, such as reference, advice, and cultural mediation.

A more recent contribution by Miguelez (2014) exploits the information on inventors’ nationality contained in PCT patent applications filed at the USPTO (more in the following section). The author estimates the impact of foreign inventors on the extent of international technological collaborations between origin and destination countries, as measured by co-patenting activity. Findings suggest a positive and significant impact for all countries of origin, that is not only for the largest ones, such as China and India.

2.3 Methodological issues

2.3.1 Name disambiguation and ethnic classification of inventors

The importance assumed by inventor data in the geography of innovation literature has pushed several scholars to improve the quality and transparency of their data mining efforts, and to discuss how their methodological choices may affect research results (Lai et al., 2011; Martínez et al., 2013; Marx et al., 2009; Pezzoni et al., 2012; 8 Notice also that regressions do not consider Western Europe and Japan as countries of origin, in order to avoid reverse causality problems: in the case of such advanced economies, it could be the case that ethnic patents in the US grow as a consequence (and not as the cause) of home technical progress, with the country of origin’s multinationals finally expanding into the US.
Raffo and Lhuillery, 2009). We sum up here some previously unexplored implications for studies on the localization and “ethnicity” of citations (for more details, see Appendix 1). 9

Ideally, a good disambiguation algorithm would minimize both “false negatives” (maximise “recall”) and “false positive” (maximise “precision”)10. Unfortunately, a trade-off exists between the two objectives, which requires making choices based on the consequences of each type of error for the subsequent analysis.

This has two consequences for the analysis of ethnic citations, based upon the linguistic analysis of inventors’ names/surnames:

1) High precision/Low recall algorithms lead to underestimating the number of personal self-citations and overestimating that of co-ethnic citations.

2) When applied to inventor sets from different countries of origin, the same matching rules may return different results in terms of precision and recall

So far, studies on migration and innovation have not adequately tackled these issues. Kerr (2007) and extensions make use of non-disambiguated inventor data (the NBER dataset), and no attempts are made to estimate the productivity of inventors, nor of the number of returnees. As for Agrawal et al. (2008, 2011) and Almeida et al. (2010), they do not provide details on the disambiguation techniques they have used, while Alnuaimi et al. (2012) apply a “perfect matching” techniques. The latter implies that only inventors with exactly the same name and surname are considered as the same person, without further checks (which in principle works as a high precision / low recall algorithm, but still can suffer of a false negative problem, due to the presence of homonyms).

Finally, we notice that precision and recall issues also appear at the ethnic classification stage, that is at the stage when inventors are assigned to a country of origin, based on their names/surnames. We discuss this matter in Appendix 3. Differently from name disambiguation, most of the studies reviewed above discuss openly this methodological issue, and decide to go for maximizing precision. For example, Agrawal et al. (2008) identify Indian inventors based on a very narrow list of Indian surnames, which are highly frequent in both India and validated by experts as indicative of recent migration status. This implies a tendency to limit the definition of ethnic inventor to first-generation migrants, which in turn hides the assumption that the strength of ethnic ties weakens with time. While making sense, the assumption is not very precise about the generational timing of the decay (at which generation do ethnic ties dissolve?) and does not consider the possibility of “ethnic revival” phenomena (as when home countries’ government try to revive contacts with their diasporas; or second- and third-generation migrants are affected by identitarian politics).

2.3.2 Harmonization of company names and identification of business groups

The JTH test as well as the studies on migration and international knowledge transfer identify company self-citations in order to exclude them from the analysis. Yet, they are silent on the methodology followed in order to uniquely identify companies, and do not mention the issues of business groups. This is in contrast with recent concerted efforts to harmonize company names as found on patent data, and assign unique identifiers to applicant (Du Plessis et al., 2009; Peeters et al., 2010; Thoma et al., 2010).

Assigning unique identifiers to patent applicants on the basis of names as reported on patent applications (without any matching of similar names) can be equated to a high precision/low recall technique (as discussed above for the case of inventors). As such, when applied to localization studies, it leads to underestimating self-citations and

---

9 The wave of interest for disambiguated inventor data has produced several open access inventor datasets. Two of them are: (i) the EP-INV dataset, originally developed for the identification of academic inventors, but comprising all inventors of patent applications filed at the European Patent Office from 1978 to around 2010 (http://www.esf-ape-inv.eu/index.php?page=3#EP-INV); and (ii) the US Patent Inventor Database, developed by Lee Fleming and associates, which contains USPTO data (http://dvn.iq.harvard.edu/dvn/dv/patent).

10 Precision and recall rates are measured as follows: 

\[
\text{Precision} = \frac{tp}{tp + fp}; \quad \text{Recall} = \frac{tp}{tp + fn}
\]

where: \(tp (fp) = \text{number of true (false)positives} \); \(tn (fn) = \text{number of true (false)negatives}\)
overestimating the co-location of knowledge flows. When applied to the analysis of international knowledge flows, it underplays the role of multinationals as carriers of knowledge towards migrants’ home countries, which in itself is an issue of substantive interest (for example, in relation to the debate concerning the strengthening of intellectual property rights in developing countries as a means for attracting foreign direct investments, and the knowledge assets that come with them; see Fink and Maskus, 2005).

Things are further complicated by the presence of business groups, which may grow by acquiring (foreign) companies and their knowledge assets. In this case, a cross-company citation may occur within the same business group, and to verify this we would need information both on the relationship between the companies involved, and the timing of the acquisition or merger.

3. Propositions and data

In this section we first synthetize our research questions by means of a set of empirical propositions. We then describe our dataset, including relevant information on methodology.

3.1 Research questions: diaspora and brain gain effect

We are interested in exploring the role of “ethnic ties” (social ties involving migrant inventors on the basis of their common country of origin) in the diffusion of knowledge, both at the national and at the international level. Ethnic ties between expatriates are interesting insofar they represent an instance of social bonds that are independent from common work experiences (co-inventorship). These social bonds may have been formed in the destination country (as a result of homophily in the choice of acquaintances and friends, possibly mediated by migrants’ associations and clubs) or inherited from the home country (as when fellow students, fellow workers, or relatives move jointly or sequentially to the same destination country). In both cases, they represent an instance of vitality and relevance of a community of expatriates, to which we will refer as a diaspora. We state a “diaspora effect” to exist when foreign inventors of the same ethnic group and active in the same country of destination (in our case, the US) have a higher propensity to cite one another’s patents, as opposed to patents by other inventors, other things being equal and excluding self-citations at the company level.

Ethnic ties may play a role at the international level, too, as when the diaspora effect is at work also to the benefit of the migrants’ home country. But they can also be indicative of the activity of multinational firms, who may send abroad some of their home country’s inventors to work in foreign branches or controlled firms, and then having their patents cited by home colleagues. Finally, they may include returnee inventors’ self-citations, which double as company self-citations when returnees are multinationals’ inventors heading back home after a spell abroad.

We first lump together all these channels under the generic heading of “brain gain”. More precisely, we state a “brain gain” effect to exist when the same foreign inventors’ patents are also disproportionately cited by inventors in the foreign inventors’ countries of origin.

We test for the existence of both the diaspora and the brain gain effect by applying the JTH methodology described in section 2. We focus on the patent applications filed by US-resident inventors coming from five Asian countries (China, India, Iran, Japan, and South Korea) and as many European ones (France, Germany, Italy, Poland, and Russia). We first run a test of proportions, which compares the share of “co-ethnic” cited-citing patent pairs with that of a control sample. We then move on to explore the determinants of a patent citation to occur by estimating the simple model:

\[
\text{Prob(citation)} = f\left(\text{co-ethnicity, spatial proximity, social proximity}\right)
\]  

(1)
In the case of the diaspora effect, the regression approach will let us comparing the relative weight of spatial, social (co-inventorship), and ethnic proximity in determining knowledge flows, as well as their interaction effects. In the case of the brain drain effect, we will chiefly interested into assessing the relative weight of ethnic citations generated by diasporas, multinationals, and returnees.

3.2 Data

3.2.1 Patent and inventor data

Our data results from matching names and surnames of inventors in the EP-INV inventor database with information on their countries of origin obtained by Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR).

The EP-INV inventor database contains information on uniquely identified inventors listed on patent applications filed at the European Patent Office (EPO). For short, we will often refer to patent applications simply as “patents”, whether granted or not.

EP-INV contains information dating back to the opening of EPO (1978) and it is continuously updated with raw data coming from PatStat, the Worldwide Patent Statistical Database published regularly by the European Patent Office. At the moment of writing this paper, EP-INV was updated to the October 2013 release of PatStat. Information on inventors includes their home address (sometimes replaced by the patent applicant's address), as harmonized and linked to Eurostat and/or OECD territorial units (NUTS3 and TL3, respectively) by RegPat, a OECD product also derived by PatStat.

Disambiguation for EP-INV is performed by making use of Massacrator 2.0, a 3-step algorithm described at length by Pezzoni et al. (2012). Massacrator 2.0’s matches inventors on the basis on the edit distances between all tokens comprised in the inventors’ name-and-surname text strings. Massacrator 2.0 does not produce a unique dataset, but several ones, each of which is calibrated against a benchmark dataset in order to return a different combination of precision and recall. Here we use a “balanced” calibration, that returns a precision rate of 88%, and a recall of 68%, when tested against a benchmark of French inventors. We also use a variation of the same calibration (“balcit”), which considers as positive cases (that is, the same person) all matched inventors whose patents are linked by at least one citation. This second calibration allows for higher recall, and directly address the problem of over-estimation of ethnic citations discussed in section 2.3.1. For all citing patents entering our final sample, we also check manually the inventors whose names and surnames are at edit distances (Levenstein distance) lower than 4.

The IBM-GNR system is a commercial product that performs various tasks, including the association of names and surnames to one or (more often) several “countries of association” (from now on: CoAs). This association originates from a database produced by US immigration authorities in the first half of the 1990s, which registered all names and surnames of all foreign citizens entering the US, along with their nationality, for a total of around 750,000 full names. In addition, variants of registered names and surnames are considered, according to country-
sensitive orthographic and abbreviation rules. As the original dataset included only non-US citizens, the US itself is never listed among the possible CoAs. 14

When fed with either a name or a surname or both, IBM-GNR returns a list of CoAs and two scores of interest:
- “frequency”, which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA;
- “significance”, which approximates the frequency distribution of the name or surname across all CoA. 15

We treat this information by means of an original algorithm (named Ethnic-INV) that we describe in Appendix 2, along with some descriptive statistics. Here it suffices to sum up the logic of the algorithm, which is as follows. Consider inventor $i$ resident in country $r$, whose name-surname combination is associated to a vector of countries [CoA], which also includes $r$. Consider also a list of Countries of Origin [CoO] of interest (in our case, the 10 countries listed in section 3.1).

The algorithm first checks whether any country $c$ included in [CoO] is also present in [CoA]. If this is the case it decides whether the frequency and significance associated to $c$ are large enough, both in absolute terms and compared to those associated to $r$, to select $c$ as inventor $i$’s Country of Origin. In case country $c$ is not included in [CoO] or it is not selected, then $i$’s Country of Origin is left unspecified (which means it may either coincide with $r$, or be one not included in the [CoO] list of interest).

Finally, we calibrate the algorithm by means of a benchmark dataset of PCT applications, which contains information on the inventors’ nationality (Miguelez and Fink, 2013). Calibration consists in altering the algorithm’s parameter in order to obtain different combinations of precision and recall. In what follows we use a “high recall” calibration, one that minimizes false negatives (it tries to avoid mistaking ethnic inventors for non-ethnic), but allows for many false positives (it mistakes many non-ethnic inventors for ethnic ones). This may introduce a negative bias in our estimates of the co-ethnicity parameter in equation (1), to the extent that, conditional on co-ethnicity to positively affect the probability of citations, a loose measurement of the latter may lead us to treat as co-ethnic two inventors who are not so, and do not cite each other’s patents. 16

Figures A3.1-2 in Appendix 3 report, respectively, the share of foreign inventors and of foreign inventors’ patent applications in the US, from 1980 to 2010, for the 10 CoO of interest (as calculated with the “high recall” calibration; patents assigned to one or another CoO as long as they include at least one of inventor from such CoO). The observed trends are close to those reported by (Kerr, 2007), with the only exception of Indian inventors’ patents in the 2000s, for which Kerr observes a decline and we do not. As for values, they are in the same order of magnitude, with our estimates for Chinese inventors’ patent share being around 1 point smaller than Kerr’s, and those for Indians’ share 1 point higher.

---

14 Information on IBM-GNR reported here comes from IBM online documentation (http://pic.dhe.ibm.com/infocenter/gnrgnm/v4r2mo/index.jsp?topic=%2Fcom.ibm.ils.gnrm.Overview.doc%2Ftopics%2FXmlr_gnrm_con_gnrm_overview.html; last visit: 4/4/2013) as well as: Patman (2010) and Nerenberg and Williams (2012). E-mail and phone exchanges with IBM staff were also decisive to facilitate our understanding. Still, being IBM-GNR a commercial product partly covered by trade secrets, we did not have entire access to its algorithms and we had to reconstruct them by deduction. For an application to a research topic close to ours, see Jeppesen and Lakhani (2010).

15 For example, an extremely common Vietnamese surname such as Nguyen will be associated both to Vietnam and to France, which hosts a significant Vietnamese minority; but in Vietnam it will get a frequency value of 90, while in France it will get only, say, 50, the Vietnamese being just a small percentage of the population. When it comes to significance, the highest percentage of inventors names Nguyen will be found in Vietnam (say 80), followed by France (which has been historically the most important destination countries of Vietnamese migrants besides the US), and several Asian countries, with much smaller values.

16 Going for a “high precision” calibration would avoid this problem, but it would also impose a considerable loss in terms of number of observations, which could affect negatively the significance of our estimates. In future versions of the paper, however, we will check the robustness of our results by running our regression exercise also on the “high precision” calibration.
As for the geographic distribution of ethnic inventors, figures A3.13-12 in Appendix 3 report the “Location Quotient” (LQ) of inventors from each CoO of interest, across all states in the US. The LQ is modelled upon the classic “relative specialization index” in international trade, varies from -1 to 1, and is defined as follows:

\[
LQ_{CoO,State} = \frac{Share_{CoO,State} - 1}{Share_{CoO,State} + 1}
\]

where:

\[
Share_{CoO,State} = \frac{Nr \text{ inventors}_{CoO,State}/Nr \text{ inventors}_{US}}{Nr \text{ inventors}_{State}/Nr \text{ inventors}_{US}}
\]

High (close to 1) values of standardized \(LQ_{CoO,State}\) indicate that the observed State host a higher-than-average share of foreign inventors from the CoO of interest. Visual inspection of maps confirms that our algorithms work well on Asian CoO, whose inventors are concentrated in US technological powerhouses (such as California, for all CoO; and Texas, for Indians). On the contrary, it may bring in too many late descendants of immigrants from European CoO (as with Polish and German inventors, whose LQ is high for several states in the Midwestern and the Great Lakes’ area).\(^{17}\)

3.2.2 Sampling for the JTH exercise and regressions

Following Agrawal et al. (2008) we focus on patents by ethnic inventors in the US, but do not follow their inventor-based sampling scheme. Rather, we stick more closely to the original patent-based JTH sampling scheme, as adapted by Breschi and Lissoni (2009) to EPO patents and to the necessity to measure social distances (distances on the network of inventors). Compared to the original JTH sampling scheme, we also control more accurately for the technological classification of patents (as suggested by Thompson and Fox-Kean, 2005).\(^{18}\)

Subject to the “high recall” calibration of the Ethnic-Inv algorithm, we select all EPO patent applications from the EP-INV database, with priority years comprised between 1990 and 2010, and at least one inventor with residence in the US, but a CoO included among the ten of our interest. We then retain only the applications that have received at least (either directly, or indirectly, via their patent family) one forward citation from another EPO patent application (either directly, or indirectly, via its patent family).\(^{19}\)

Notice that the same cited patent enter our sample as many times as the number of citations it receives. The same applies to each citing patent that cites more than one cited patent. This will require correcting for non-independence of errors when conducting our econometric exercises.

We then proceed to collect information on:

- the citing and cited patents’ applicants (as harmonized by the EEE-PPAT project and published with the October 2013 release of PatStat, supplemented by manual checks)\(^{20}\)
- the inventors’ identity, as from the EP-INV database; in first instance, we make use of EP-INV “balcit” unique identifiers (see above)

---

\(^{17}\) As for Italians and Russians, they exhibit a high LQ in New York state, which host both recent migrants and the descendants of early ones), while Louisiana is among the states with LQ>1 for French inventors, which also sounds suspect.

\(^{18}\) Agrawal et al.’s (2008) sampling scheme also differs from JTH in that, once identified the cited-citing patent pairs, it proceed to build the control sample by matching on cited patents, instead of citing ones.

\(^{19}\) Several patents from the same or, more commonly, different patent offices, form a « family » whenever they share one or several priority filings (or “priorities”). Roughly speaking, the family includes all patent documents that protect the same invention, so it is good practice to measure the citation links between two patents from the same office (in our case, the EPO) by counting all citations running between the families the two patents belong to (Harhoff et al., 2003). Several definitions of family exist, of which we adopt here the simplest one: all patents in the family must share exactly the same priorities (for more definition and a discussion, see Martinez, 2011).

• the inventors’ addresses (country and, for US-residents, the BEA Economic Area), also from EP-INV
• the citing patents’ priority year and technological field (IPC group) 21

On this basis, we proceed to build two different samples, a “local” and an “international” one.

For the local sample we retain all cited-citing pairs in which the citing patent comprises among its inventors at least one US-resident, and we exclude all self-citations at the applicant level, as well as all self-citations at the inventor level, where the self-citing inventor belongs to one of the 10 CoO of interest. For each citing patent, we randomly select a control patent that satisfies the following conditions:

• it does not cite the cited patent in cited-citing pair
• it has the same priority year and is classified under the same IPC groups of the citing patent 22
• it comprises among its inventor at least one US-resident

This leaves us with 1,006,206 observations, half of which (503,103) are cited-citing pairs, the other half cited-control pairs. The cited-citing pairs are generated by the combination of 96,800 cited patents and 216,115 citing ones, with the former receiving on average 2.2 citations. Table 1 (part 1) reports details by CoO. As expected, more than half the observations come from the two largest CoO, China and India. The only European country in the same order of magnitude is Germany, which however may include patents by Swiss and Austrian inventors, due to imprecisions of our algorithm.

Intuitively, the local sample includes all the citations addressed to “ethnic” patents by US-resident inventors. Were we to find that ethnicity affecting the observed citation patterns, this would point at the existence of a diaspora effect (social ties among migrants in the same destination country, from the same country of origin).

As for the “international” sample we retain all cited-citing pairs in which the citing patent has no US-resident inventors. For each citing patent, we randomly select a control patent that satisfies the following conditions:

• it does not cite the cited patent in cited-citing pair
• it has the same priority year and is classified under the same IPC groups of the citing patent
• it does not comprises among its inventor any US-resident

This leaves us with 889,832 observations, half of which (444,916) are cited-citing pairs, the other half cited-control pairs. The cited-citing pairs are generated by the combination of 106,237 cited patents and 272,697 citing ones, with the former receiving on average 2.6 citations. Table 1 (part 2) shows that the distribution of observations by CoO of cited patents’ inventors is the same as that for the local sample.

Notice that, differently than for the local sample, we retain all self-citations at the applicant level, as well as all self-citations at the inventor level. We do so because the international sample is meant to investigate international, ethnic-mediated knowledge flows. These may include intra-company exchanges, as well as knowledge transfers by returnee inventors, which explain why the average number of citations is higher.

When running a test of proportions, the cited-citing pairs and the cited-control pairs in both the local and the international samples will be treated as different subsamples (in a “cases vs controls” setting). In the regression setting, they will be “stacked” and flagged as different by means of the binary variable Citation (=1 for cited-citing pairs, =0 for cited-control pairs), which will then be our dependent variable in the econometric exercises.

21 IPC is the International Patent Classification, which is maintained and regularly update by the World Intellectual Property Organization (WIPO). We use IPC version 8. IPC groups are the second finest level of aggregation. For details: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf (last visited, May 2014)

22 Notice that the same patent may be assigned to several IPC groups. Therefore our matching criteria require the citing patent and its control to be classified under the same number of IPC groups, and to share them all
Table 1. Local and international samples: nr of patents, pairs, and observations; by country of origin of cited patents’ inventors

<table>
<thead>
<tr>
<th>Country</th>
<th>Nr cited patents</th>
<th>Citing patents</th>
<th>Cited-citing pairs</th>
<th>Obs (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nr</td>
<td>%</td>
<td>Nr</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Local sample (citations from within the US)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>29632</td>
<td>25.3</td>
<td>83268</td>
<td>20.8</td>
</tr>
<tr>
<td>Germany</td>
<td>19100</td>
<td>16.3</td>
<td>71696</td>
<td>17.9</td>
</tr>
<tr>
<td>France</td>
<td>7534</td>
<td>6.4</td>
<td>30584</td>
<td>7.6</td>
</tr>
<tr>
<td>India</td>
<td>35686</td>
<td>30.5</td>
<td>108726</td>
<td>27.1</td>
</tr>
<tr>
<td>Iran</td>
<td>3170</td>
<td>2.7</td>
<td>13992</td>
<td>3.5</td>
</tr>
<tr>
<td>Italy</td>
<td>4560</td>
<td>3.9</td>
<td>21172</td>
<td>5.3</td>
</tr>
<tr>
<td>Japan</td>
<td>5351</td>
<td>4.6</td>
<td>22781</td>
<td>5.7</td>
</tr>
<tr>
<td>Korea</td>
<td>5582</td>
<td>4.8</td>
<td>23111</td>
<td>5.8</td>
</tr>
<tr>
<td>Poland</td>
<td>1901</td>
<td>1.6</td>
<td>8142</td>
<td>2.0</td>
</tr>
<tr>
<td>Russia</td>
<td>4506</td>
<td>3.9</td>
<td>17017</td>
<td>4.2</td>
</tr>
<tr>
<td>Total (1)</td>
<td>117025</td>
<td>100</td>
<td>400489</td>
<td>100</td>
</tr>
<tr>
<td>Total (2)</td>
<td>96800</td>
<td>100</td>
<td>216115</td>
<td>100</td>
</tr>
<tr>
<td>2. International sample (citations from outside the US)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>31777</td>
<td>24.8</td>
<td>91356</td>
<td>21.9</td>
</tr>
<tr>
<td>Germany</td>
<td>21767</td>
<td>17.0</td>
<td>74843</td>
<td>A 17.9</td>
</tr>
<tr>
<td>France</td>
<td>8353</td>
<td>6.5</td>
<td>30202</td>
<td>7.2</td>
</tr>
<tr>
<td>India</td>
<td>38484</td>
<td>30.0</td>
<td>11871</td>
<td>28.3</td>
</tr>
<tr>
<td>Iran</td>
<td>3356</td>
<td>2.6</td>
<td>12395</td>
<td>3.0</td>
</tr>
<tr>
<td>Italy</td>
<td>5087</td>
<td>4.0</td>
<td>20481</td>
<td>4.9</td>
</tr>
<tr>
<td>Japan</td>
<td>6263</td>
<td>4.9</td>
<td>24011</td>
<td>5.7</td>
</tr>
<tr>
<td>Korea</td>
<td>6041</td>
<td>4.7</td>
<td>21830</td>
<td>5.2</td>
</tr>
<tr>
<td>Poland</td>
<td>2115</td>
<td>1.6</td>
<td>7613</td>
<td>1.8</td>
</tr>
<tr>
<td>Russia</td>
<td>4975</td>
<td>3.9</td>
<td>16839</td>
<td>4.0</td>
</tr>
<tr>
<td>Total (1)</td>
<td>128218</td>
<td>100</td>
<td>417741</td>
<td>100</td>
</tr>
<tr>
<td>Total (2)</td>
<td>106237</td>
<td>100</td>
<td>272679</td>
<td>100</td>
</tr>
</tbody>
</table>

(1) Total = sum of observations by country of origin (same patent may be recorded under >1 country)
(2) Total = sum of distinct observations
(3) Nr observations per country = Nr cited-citing pairs * 2

For all patent pairs in the two samples, we produce the following dummy variables, which will enter as independent variables in the regressions:

1. Co-ethnicity: =1 iff at least one inventor in the cited patent and one inventor in the citing (control) one are from the same CoO
2. Social distance S (with S=0,1,2,>3,+∞): =1 iff the minimum geodesic distance between cited patent and the citing (control) is equal to S. Formally: \( S = \min \{ S_{ij} \} \) with \( S_{ij} \) = geodesic distance between inventor \( i \) (\( i = 1 \ldots I \)) on the cited patent and inventor \( j \) (\( j = 1 \ldots J \)) on the citing (control) one, as calculated on the entire network of inventors, for all inventors on the cited and the citing (control) patents. Notice that: for \( i = j \rightarrow S = 0 \); if all \( i \) and all \( j \) belong to disconnected network components: \( S = +\infty \). For each year \( t \) we calculate a different network of inventors, based on co-inventorship patterns of all patents with priority years (t-1, t-5). 23

For patent pairs in the local sample we also calculate:
3. Co-location: =1 iff at least one inventor in the cited patent and one inventor in the citing (control) are located in the same BEA economic area.

For patent pairs in the international sample we also calculate:

---

23 This amounts to assuming that social ties generated by co-inventorship decay after 5 years, unless renewed by further collaborations. For more details, see Breschi and Lissoni (2009).
4. **Same country**: =1 iff the country of residence of the citing (control) patent’s inventor is the same as the CoO of the cited patent’s inventor.

5. **Same company**: =1 iff applicants of the cited and the citing (control) patents are the same.

6. **Returnee**: =1 iff the inventor of the cited and the citing (control) patents are the same (notice that this implies Social distance 0 = 1)

Table 2 reports the descriptive statistics for all variables in both samples; for details by country, see tables A3.13-22 in the appendix.

### Table 2. Local and international samples: descriptive statistics

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Local sample (citations from within the US)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citation</td>
<td>1211154</td>
<td>0.500</td>
<td>0.500</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Co-ethnicity</td>
<td>1211154</td>
<td>0.120</td>
<td>0.325</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 0</td>
<td>1211154</td>
<td>0.013</td>
<td>0.114</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 1</td>
<td>1211154</td>
<td>0.012</td>
<td>0.109</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 2</td>
<td>1211154</td>
<td>0.008</td>
<td>0.089</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 3</td>
<td>1211154</td>
<td>0.009</td>
<td>0.093</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance &gt;3</td>
<td>1211154</td>
<td>0.236</td>
<td>0.425</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance +∞</td>
<td>1211154</td>
<td>0.722</td>
<td>0.448</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Co-location</td>
<td>1211154</td>
<td>0.172</td>
<td>0.377</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>2. International sample (citations from outside the US)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citation</td>
<td>1084120</td>
<td>0.500</td>
<td>0.500</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Co-ethnicity</td>
<td>1084120</td>
<td>0.081</td>
<td>0.272</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 0</td>
<td>1084120</td>
<td>0.004</td>
<td>0.063</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 1</td>
<td>1084120</td>
<td>0.005</td>
<td>0.072</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 2</td>
<td>1084120</td>
<td>0.004</td>
<td>0.066</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance 3</td>
<td>1084120</td>
<td>0.005</td>
<td>0.068</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance &gt;3</td>
<td>1084120</td>
<td>0.200</td>
<td>0.400</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social distance +∞</td>
<td>1084120</td>
<td>0.781</td>
<td>0.413</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Same country</td>
<td>1084120</td>
<td>0.085</td>
<td>0.279</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Same company</td>
<td>1084120</td>
<td>0.024</td>
<td>0.152</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Returnee</td>
<td>1084120</td>
<td>0.0005</td>
<td>0.022</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### 4. Results and discussion

#### 4.1 Within-US knowledge flows

We test the existence of a diaspora effect on data from the local sample (citations within the US), first by running some simple tests of proportions, then by means of a logit regression. We run both the tests and the regression separately by CoO of the cited patents’ inventors.

Table 3 reports the results of two simple tests of proportions, the former comparing the co-location rate of cited-citing patent pairs to that of control pairs, the latter doing the same for the co-ethnicity. We interpret any positive evidence on the co-ethnicity of citations as indicative of the existence of a diaspora effect, for the CoO considered. As for co-location, we are interested in checking whether our “ethnic” data reproduce the JTH basic results (which were obtained for all US-resident inventors), and in assessing to what extent they owe to social and ethnic ties.
Table 3. Co-location and co-ethnicity of citations: test of proportions, by CoO

<table>
<thead>
<tr>
<th>CoO</th>
<th>citing</th>
<th>controls</th>
<th>z</th>
<th>citing</th>
<th>controls</th>
<th>z</th>
<th>obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>22.3</td>
<td>13.2</td>
<td>64.19*</td>
<td>24.5</td>
<td>17.9</td>
<td>43.53*</td>
<td>145902</td>
</tr>
<tr>
<td>Germany</td>
<td>20.9</td>
<td>12.1</td>
<td>53.84*</td>
<td>7.6</td>
<td>7</td>
<td>5.56*</td>
<td>102929</td>
</tr>
<tr>
<td>France</td>
<td>22</td>
<td>13.5</td>
<td>30.86*</td>
<td>3.9</td>
<td>3.5</td>
<td>3.1*</td>
<td>38519</td>
</tr>
<tr>
<td>India</td>
<td>21.3</td>
<td>13</td>
<td>67.56*</td>
<td>17.3</td>
<td>14.6</td>
<td>22.79*</td>
<td>186563</td>
</tr>
<tr>
<td>Iran</td>
<td>24.4</td>
<td>14.3</td>
<td>23.37*</td>
<td>1.7</td>
<td>1.2</td>
<td>3.74*</td>
<td>16564</td>
</tr>
<tr>
<td>Italy</td>
<td>19.4</td>
<td>11.7</td>
<td>24.35*</td>
<td>2</td>
<td>1.9</td>
<td>1.54</td>
<td>26584</td>
</tr>
<tr>
<td>Japan</td>
<td>20.9</td>
<td>13</td>
<td>25.01*</td>
<td>3</td>
<td>2.4</td>
<td>4.2*</td>
<td>28117</td>
</tr>
<tr>
<td>Korea</td>
<td>20.7</td>
<td>13</td>
<td>24.9*</td>
<td>3.2</td>
<td>2.5</td>
<td>5.04*</td>
<td>29728</td>
</tr>
<tr>
<td>Poland</td>
<td>19.8</td>
<td>11.2</td>
<td>16.58*</td>
<td>0.6</td>
<td>0.8</td>
<td>-1.03</td>
<td>9539</td>
</tr>
<tr>
<td>Russia</td>
<td>22.7</td>
<td>12.5</td>
<td>27.52*</td>
<td>3.4</td>
<td>2.4</td>
<td>5.75*</td>
<td>21132</td>
</tr>
</tbody>
</table>

* significance at 99%

The co-location test replicates, to a large extent, JTH’s classic results, with percentages that do not vary sensibly across CoO. As for the co-ethnicity test, this is positive and significant for all CoO, with the only exception of Poland, but it hides important cross-country differences. In particular, it seems to suggest that diaspora effects may be the strongest for China, India, Russia, and South Korea. The former two record the highest percentages of co-ethnic citing-cited and control-cited pairs, which is explained by the large number of Chinese and Indian inventors in the sample. They also report the largest difference between the co-ethnicity percentages of cited-citing and cited-control pairs (6.6 and 2.7, respectively). As for Russia, it records the third largest difference in absolute value, and the largest ratio (1.0 and 1.42, respectively; the latter *ex aequo* with Iran). South Korea follows in fourth position (for both difference and ratio).

Tables 4a-c report the results of three different specifications of model (1), which we estimate by means of a logit regression. The first specification reproduces Agrawal et al.’s (2008) basic exercise for Indian inventors in the US; the second introduces social distances between inventors; the latter allows for interactions between such distances and ethnic ties. All tables report the estimated parameters; robust standard errors are calculated by clustering on patent pairs.

Table 4a reproduces Agrawal et al.’s (2008) findings, most notably for China and India: co-ethnicity affects positively the probability to observe a citation link between two patents, but its marginal effect is smaller than that of co-location. Besides, the interaction term between co-ethnicity and co-location is negative, which suggests a substitution effect between the two. As for other countries, we never find the interaction term to be significant; and we find co-ethnicity to matter for Germany, Japan, Korean, and Russia, but not for France, Italy, and Poland.

Table 4b reproduces Breschi and Lissoni’s (2009) findings, for what concerns the role of social distances between inventors. First, we observe the estimated parameters for the latter to be always negative and significant, as well as increasing in absolute value when moving from Social distance 1 to Social Distance +∞: being Social distance 0 the reference cases, this suggest that the probability to observe a citation between two patents diminishes with the increase of social distance. In addition, when controlling for social distance the estimated parameters for Co-location get smaller, which confirms that, when failing to control for social distance, estimates of the importance of co-location are positively biased. As for Co-ethnicity, the sign and significance of both its parameter and of the interaction term with Co-location do not change much.
Finally, table 4c introduces two interaction terms for social distance and co-ethnicity. In order to avoid having too many of them, we reduce the dummy variables for social distance to three (instead of six): Social distance $\leq 3$, Social Distance $>3$, and Social Distance $+\infty$. We observe that, with this specification, co-ethnicity seem to matter only for China and India, as for all other CoO the estimated parameter (of both the direct and the interaction terms) appear not to be significant. As for China and India, the parameters for Co-Ethnicity becomes negative, while the interaction terms with Social Distance $>3$ and Social Distance $+\infty$ are both positive and comparatively large (in absolute value). This suggests co-ethnicity and social proximity (the inverse of social proximity).
distance) to be substitutes, as already observed for co-ethnicity and spatial proximity (co-location); ethnic ties act as “shortcuts” over the network of inventors, and allow reaching otherwise distance actors. This interpretation is illustrated by figure 1, which reports the predicted probabilities to observe a citation for China and India, as a function of co-location, co-ethnicity, and social distance, based on estimates from table 4c.

Table 4c – Ethnic inventors’ patents: probability of citation from within the US, as a function of co-location, co-ethnicity, and social distance (with interactions) — Logit regression, by Country of Origin

<table>
<thead>
<tr>
<th></th>
<th>China</th>
<th>Germany</th>
<th>France</th>
<th>India</th>
<th>Iran</th>
<th>Italy</th>
<th>Japan</th>
<th>Korea</th>
<th>Poland</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-location</td>
<td>0.407***</td>
<td>0.454***</td>
<td>0.410***</td>
<td>0.422***</td>
<td>0.477***</td>
<td>0.413***</td>
<td>0.390***</td>
<td>0.353***</td>
<td>0.326***</td>
<td>0.487***</td>
</tr>
<tr>
<td></td>
<td>(0.0117)</td>
<td>(0.0126)</td>
<td>(0.0194)</td>
<td>(0.0098)</td>
<td>(0.0285)</td>
<td>(0.0243)</td>
<td>(0.0230)</td>
<td>(0.0225)</td>
<td>(0.0419)</td>
<td>(0.0268)</td>
</tr>
<tr>
<td>Co-ethnicity</td>
<td>-0.290***</td>
<td>0.0645</td>
<td>-0.213</td>
<td>-0.199***</td>
<td>0.366</td>
<td>0.171</td>
<td>0.454</td>
<td>0.0301</td>
<td>-0.459</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>(0.0530)</td>
<td>(0.105)</td>
<td>(0.0260)</td>
<td>(1.059)</td>
<td>(0.367)</td>
<td>(0.289)</td>
<td>(0.741)</td>
<td>(0.741)</td>
<td>(0.523)</td>
<td></td>
</tr>
<tr>
<td>Co-ethn*Co-loc</td>
<td>-0.097***</td>
<td>-0.0515</td>
<td>0.0505</td>
<td>-0.073***</td>
<td>0.128</td>
<td>-0.172</td>
<td>-0.0948</td>
<td>-0.101</td>
<td>-0.242</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>(0.0241)</td>
<td>(0.0463)</td>
<td>(0.0949)</td>
<td>(0.0236)</td>
<td>(0.217)</td>
<td>(0.161)</td>
<td>(0.138)</td>
<td>(0.123)</td>
<td>(0.450)</td>
<td>(0.159)</td>
</tr>
<tr>
<td>Soc. distance &gt;3</td>
<td>-1.912***</td>
<td>-1.656***</td>
<td>-1.610***</td>
<td>-1.783***</td>
<td>-1.838***</td>
<td>-1.722***</td>
<td>-1.875***</td>
<td>-1.762***</td>
<td>-1.953***</td>
<td>-1.769***</td>
</tr>
<tr>
<td></td>
<td>(0.0323)</td>
<td>(0.0357)</td>
<td>(0.0562)</td>
<td>(0.0294)</td>
<td>(0.0872)</td>
<td>(0.0734)</td>
<td>(0.0742)</td>
<td>(0.0653)</td>
<td>(0.105)</td>
<td>(0.0766)</td>
</tr>
<tr>
<td>Soc. distance +∞</td>
<td>-2.019***</td>
<td>-1.759***</td>
<td>-1.747***</td>
<td>-1.879***</td>
<td>-1.878***</td>
<td>-1.817***</td>
<td>-1.956***</td>
<td>-1.869***</td>
<td>-2.031***</td>
<td>-1.937***</td>
</tr>
<tr>
<td></td>
<td>(0.0315)</td>
<td>(0.0344)</td>
<td>(0.0542)</td>
<td>(0.0286)</td>
<td>(0.0841)</td>
<td>(0.0711)</td>
<td>(0.0724)</td>
<td>(0.0633)</td>
<td>(0.102)</td>
<td>(0.0740)</td>
</tr>
<tr>
<td>Co-ethn*Soc. Distance &gt;3</td>
<td>0.761***</td>
<td>0.00230</td>
<td>0.215</td>
<td>0.414***</td>
<td>-0.0882</td>
<td>0.0402</td>
<td>-0.158</td>
<td>0.184</td>
<td>0.551</td>
<td>-0.404</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.107)</td>
<td>(0.211)</td>
<td>(0.0606)</td>
<td>(1.068)</td>
<td>(0.380)</td>
<td>(0.396)</td>
<td>(0.293)</td>
<td>(0.737)</td>
<td>(0.530)</td>
</tr>
<tr>
<td>Co-ethn*Soc. Distance +∞</td>
<td>0.554***</td>
<td>-0.0337</td>
<td>0.258</td>
<td>0.368***</td>
<td>-0.0901</td>
<td>-0.209</td>
<td>-0.328</td>
<td>0.146</td>
<td>0.0826</td>
<td>-0.128</td>
</tr>
<tr>
<td></td>
<td>(0.0535)</td>
<td>(0.105)</td>
<td>(0.208)</td>
<td>(0.0591)</td>
<td>(1.058)</td>
<td>(0.371)</td>
<td>(0.385)</td>
<td>(0.290)</td>
<td>(0.749)</td>
<td>(0.522)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.775***</td>
<td>1.607***</td>
<td>1.588***</td>
<td>1.706***</td>
<td>1.719***</td>
<td>1.685***</td>
<td>1.819***</td>
<td>1.723***</td>
<td>1.887***</td>
<td>1.747***</td>
</tr>
<tr>
<td></td>
<td>(0.0315)</td>
<td>(0.0344)</td>
<td>(0.0542)</td>
<td>(0.0285)</td>
<td>(0.0839)</td>
<td>(0.0709)</td>
<td>(0.0723)</td>
<td>(0.0632)</td>
<td>(0.102)</td>
<td>(0.0740)</td>
</tr>
<tr>
<td>Observations</td>
<td>291,804</td>
<td>205,858</td>
<td>77,038</td>
<td>373,126</td>
<td>33,128</td>
<td>53,168</td>
<td>56,234</td>
<td>59,456</td>
<td>19,078</td>
<td>42,264</td>
</tr>
<tr>
<td>Chi-sq</td>
<td>11787</td>
<td>5730</td>
<td>2094</td>
<td>10150</td>
<td>974.1</td>
<td>1250</td>
<td>1300</td>
<td>1527</td>
<td>622.7</td>
<td>1425</td>
</tr>
<tr>
<td>LogL</td>
<td>-195749</td>
<td>-139315</td>
<td>-52525</td>
<td>-252663</td>
<td>-22348</td>
<td>-36090</td>
<td>-38124</td>
<td>-40287</td>
<td>-12819</td>
<td>-28421</td>
</tr>
<tr>
<td>Pseudo R-sq</td>
<td>0.0322</td>
<td>0.0237</td>
<td>0.0222</td>
<td>0.0231</td>
<td>0.0268</td>
<td>0.0207</td>
<td>0.0219</td>
<td>0.0224</td>
<td>0.0306</td>
<td>0.0299</td>
</tr>
</tbody>
</table>

The table reports estimated parameters (β); Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

We first observe that the marginal effect of social distance appear to dominate both that of co-location and that of co-ethnicity: when moving from Social distance ≤3 to Social Distance >3, and Social Distance +∞ the probability of citation falls drastically, and it is never compensated by the occurrence of co-location or co-ethnicity. Second, we observe that, at Social distance ≤3, the marginal effects of both co-location [move from the (0,0) to the (1,0) bar] or co-ethnicity [move from (0,0) to (0,1) bar] are rather limited; while they are more noticeable Social Distance >3 and Social Distance +∞. Second, the substitution effect between co-location and co-ethnicity is stronger for short social distances than for large ones: moving from the (0,0) to the (1,1) bar for Social distance ≤3 induces almost no increase in the probability of citation, which instead increases sensibly when the move occurs for Social Distance >3 and Social Distance +∞.
4.2 International knowledge flows

Tables 5a-c reports the results of three different specifications of a logit regression exercise based upon the international sample, where the dependent variable is the probability of a US-resident ethnic inventor’s patent to be cited by a non US-resident inventor’s patent. The sample include self-citations both at the individual level (returnee inventors’ self-citations) and at the company level.

No specification includes Returnee as a regressor. This is because, for several CoO, the latter would predict perfectly the value of the dependent variable. As for Same company, this affect positively and significantly the probability of citations in all specification, as expected. The value of its estimated coefficient is slightly affected, in the specification of table 5c, by the insertion of social distance as an additional regressor; however, no interaction term with the latter is ever significant (so we do not report specification that include interactions between the two).
All regressions concerning Iran do not pass the LogL test, and the same applies to Poland for specification in table 5c.

<table>
<thead>
<tr>
<th></th>
<th>China</th>
<th>Germany</th>
<th>France</th>
<th>India</th>
<th>Iran</th>
<th>Italy</th>
<th>Japan</th>
<th>Korea</th>
<th>Poland</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same country</td>
<td>0.138*</td>
<td>-0.194**</td>
<td>0.0276</td>
<td>0.171***</td>
<td>0.0523</td>
<td>-0.185***</td>
<td>0.212***</td>
<td>0.425**</td>
<td>0.358***</td>
<td>0.598</td>
</tr>
<tr>
<td>(0.0243)</td>
<td>(0.0101)</td>
<td>(0.0340)</td>
<td>(0.0370)</td>
<td>(1.000)</td>
<td>(0.0460)</td>
<td>(0.0183)</td>
<td>(0.0413)</td>
<td>(0.479)</td>
<td>(0.152)</td>
<td></td>
</tr>
<tr>
<td>Same company</td>
<td>1.173***</td>
<td>1.087***</td>
<td>1.236***</td>
<td>1.117***</td>
<td>1.258***</td>
<td>0.902***</td>
<td>1.332***</td>
<td>0.968***</td>
<td>1.106***</td>
<td>1.133***</td>
</tr>
<tr>
<td>(0.0365)</td>
<td>(0.0327)</td>
<td>(0.0591)</td>
<td>(0.0323)</td>
<td>(0.118)</td>
<td>(0.0796)</td>
<td>(0.0578)</td>
<td>(0.0817)</td>
<td>(0.152)</td>
<td>(0.101)</td>
<td></td>
</tr>
<tr>
<td>Social distance 1</td>
<td>-1.205***</td>
<td>-1.268***</td>
<td>-2.047***</td>
<td>-2.682***</td>
<td>-17.05</td>
<td>-1.118**</td>
<td>-0.739**</td>
<td>-1.576**</td>
<td>-0.867</td>
<td>-1.331**</td>
</tr>
<tr>
<td>(0.262)</td>
<td>(0.189)</td>
<td>(0.449)</td>
<td>(0.423)</td>
<td>(0)</td>
<td>(0.515)</td>
<td>(0.371)</td>
<td>(0.765)</td>
<td>(1.132)</td>
<td>(0.590)</td>
<td></td>
</tr>
<tr>
<td>(0.251)</td>
<td>(0.186)</td>
<td>(0.434)</td>
<td>(0.417)</td>
<td>(0.348)</td>
<td>(0.483)</td>
<td>(0.357)</td>
<td>(0.742)</td>
<td>(1.105)</td>
<td>(0.549)</td>
<td></td>
</tr>
<tr>
<td>(0.246)</td>
<td>(0.184)</td>
<td>(0.427)</td>
<td>(0.414)</td>
<td>(0.341)</td>
<td>(0.478)</td>
<td>(0.346)</td>
<td>(0.730)</td>
<td>(1.074)</td>
<td>(0.542)</td>
<td></td>
</tr>
<tr>
<td>(0.239)</td>
<td>(0.172)</td>
<td>(0.412)</td>
<td>(0.410)</td>
<td>(0.275)</td>
<td>(0.453)</td>
<td>(0.313)</td>
<td>(0.715)</td>
<td>(1.007)</td>
<td>(0.510)</td>
<td></td>
</tr>
<tr>
<td>(0.239)</td>
<td>(0.172)</td>
<td>(0.412)</td>
<td>(0.410)</td>
<td>(0.273)</td>
<td>(0.452)</td>
<td>(0.313)</td>
<td>(0.715)</td>
<td>(1.006)</td>
<td>(0.509)</td>
<td></td>
</tr>
<tr>
<td>(0.239)</td>
<td>(0.172)</td>
<td>(0.412)</td>
<td>(0.409)</td>
<td>(0.273)</td>
<td>(0.452)</td>
<td>(0.313)</td>
<td>(0.715)</td>
<td>(1.006)</td>
<td>(0.509)</td>
<td></td>
</tr>
</tbody>
</table>

The table reports estimated parameters (β); Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Specifications in table 5a and 5b differ only in the choice of one regressor, with Same country in the former being replaced by Co-ethnicity in the latter. Same country is our variable of choice for capturing the existence of ties between a migrant inventor and her home country, besides those mediated by multinational companies. We expect its sign to be positive and significant, which is the case for all Asian countries and Russia. For France and Poland the estimated coefficient is not significant, while for Germany and Italy it is even negative. When we replace Same country with Co-ethnicity we always get positive and significant coefficients, with the exception of Poland; for all countries (except Japan, Korea, and Russia), the value of the coefficient increases considerably.

This contrast between the two specifications may have two explanations, one concerning our data’s quality, the other being possibly more substantive. The first explanation is as follows. It may be the case that many US-resident inventors that we identify as Germans or Italians are migrants not just from Germany and Italy, but from, respectively, other German-speaking countries (such as Switzerland and Austria) and several countries with sizeable communities of Italian migrants’ descendants (such as Brazil, Argentina, or a few Western European countries). This implies that many citations to what we identify as German or Italian inventors’ patents are in fact citations to, say, Swiss or Brazilian inventors’ patents, which do not come from Germany or Italy (for which Same country = 1), but from Switzerland or Brazil (Same country = 0). The same may apply to France and, to a more limited extent, China and India. In the case of China, several citations to Chinese inventors in our sample may come from Honk Kong and Taiwan; in the case of India from the several Commonwealth countries (which host a sizeable community of highly educated Indian migrants or British citizens of Indian descent).
Table 5b – Ethnic inventors’ patents: probability of international citation, as a function of co-location, co-ethnicity, and social distance -- Logit regression, by Country of Origin

<table>
<thead>
<tr>
<th>Co-ethnicity</th>
<th>China</th>
<th>Germany</th>
<th>France</th>
<th>India</th>
<th>Iran</th>
<th>Italy</th>
<th>Japan</th>
<th>Korea</th>
<th>Poland</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-motion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social distance &gt;3</td>
<td>1.16**</td>
<td>1.24**</td>
<td>1.14**</td>
<td>1.14**</td>
<td>1.16**</td>
<td>0.87*</td>
<td>1.24**</td>
<td>0.99**</td>
<td>0.94**</td>
<td>0.91**</td>
</tr>
<tr>
<td>Same company</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social distance 2</td>
<td>2.33**</td>
<td>1.95**</td>
<td>2.77**</td>
<td>3.60**</td>
<td>17.62**</td>
<td>1.96**</td>
<td>1.66**</td>
<td>3.03**</td>
<td>1.85*</td>
<td>2.69**</td>
</tr>
<tr>
<td>Social distance 3</td>
<td>3.07**</td>
<td>2.64**</td>
<td>3.52**</td>
<td>4.34**</td>
<td>18.41**</td>
<td>3.18**</td>
<td>2.64**</td>
<td>3.91**</td>
<td>2.97**</td>
<td></td>
</tr>
<tr>
<td>Soc. distance &gt;3</td>
<td>3.09**</td>
<td>2.64**</td>
<td>3.52**</td>
<td>4.34**</td>
<td>18.41**</td>
<td>3.18**</td>
<td>2.64**</td>
<td>3.91**</td>
<td>2.97**</td>
<td></td>
</tr>
<tr>
<td>Soc. distance +∞</td>
<td>3.25**</td>
<td>2.70**</td>
<td>3.67**</td>
<td>4.46**</td>
<td>18.50**</td>
<td>3.31**</td>
<td>2.69**</td>
<td>4.09**</td>
<td>3.51**</td>
<td>3.87**</td>
</tr>
<tr>
<td>Constant</td>
<td>3.16**</td>
<td>2.57**</td>
<td>3.53**</td>
<td>4.40**</td>
<td>18.45**</td>
<td>3.24**</td>
<td>2.54**</td>
<td>4.01**</td>
<td>3.45**</td>
<td>3.11**</td>
</tr>
</tbody>
</table>

The table reports estimated parameters (β); Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 5c – Ethnic inventors’ patents: probability of international citation, as a function of co-location, co-ethnicity, and social distance (with interactions) -- Logit regression, by Country of Origin

<table>
<thead>
<tr>
<th>Co-ethnicity</th>
<th>China</th>
<th>Germany</th>
<th>France</th>
<th>India</th>
<th>Iran</th>
<th>Italy</th>
<th>Japan</th>
<th>Korea</th>
<th>Poland</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-motion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social distance &gt;3</td>
<td>1.16**</td>
<td>1.24**</td>
<td>1.14**</td>
<td>1.14**</td>
<td>1.16**</td>
<td>0.87*</td>
<td>1.24**</td>
<td>0.99**</td>
<td>0.94**</td>
<td>0.91**</td>
</tr>
<tr>
<td>Same company</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social distance 2</td>
<td>2.33**</td>
<td>1.95**</td>
<td>2.77**</td>
<td>3.60**</td>
<td>17.62**</td>
<td>1.96**</td>
<td>1.66**</td>
<td>3.03**</td>
<td>1.85*</td>
<td>2.69**</td>
</tr>
<tr>
<td>Social distance 3</td>
<td>3.07**</td>
<td>2.64**</td>
<td>3.52**</td>
<td>4.34**</td>
<td>18.41**</td>
<td>3.18**</td>
<td>2.64**</td>
<td>3.91**</td>
<td>2.97**</td>
<td></td>
</tr>
<tr>
<td>Soc. distance &gt;3</td>
<td>3.09**</td>
<td>2.64**</td>
<td>3.52**</td>
<td>4.34**</td>
<td>18.41**</td>
<td>3.18**</td>
<td>2.64**</td>
<td>3.91**</td>
<td>2.97**</td>
<td></td>
</tr>
<tr>
<td>Soc. distance +∞</td>
<td>3.25**</td>
<td>2.70**</td>
<td>3.67**</td>
<td>4.46**</td>
<td>18.50**</td>
<td>3.31**</td>
<td>2.69**</td>
<td>4.09**</td>
<td>3.51**</td>
<td>3.87**</td>
</tr>
<tr>
<td>Constant</td>
<td>3.16**</td>
<td>2.57**</td>
<td>3.53**</td>
<td>4.40**</td>
<td>18.45**</td>
<td>3.24**</td>
<td>2.54**</td>
<td>4.01**</td>
<td>3.45**</td>
<td>3.11**</td>
</tr>
</tbody>
</table>

The table reports estimated parameters (β); Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1
As for the second explanation, this points to the possible existence of “international diaspora” effects, namely ethnic ties that go beyond the boundaries of one single migrants’ destination country (in our case, the US). That is, especially for the increase in the value of the co-efficient for China and India, our results could be indicative of ethnic ties that span over several destination countries. At this stage of our research, however, this remark is only speculative.

Coming to social distance, the dummies that represent it are always significant and have the expected sign, with their absolute value increasing with distance in all regressions. When we introduce an interaction term with co-ethnicity (table 5-c), the latter’s parameters, as well as the parameters of the interaction terms, become not significant for Italy, Japan, Korea, and Russia. Based on the specification with interactions, figure 2 reports the estimated probability of international citations for India, as a function of social distance, co-ethnicity, and the patent ownership by the same company. We notice that the marginal effect (increase in probability) of co-ethnicity is the largest for large social distances, and for patents that do not belong to the same company. Similarly to what found for co-location in the case of within-US citations, this suggest that co-ethnicity may kick in as a substitute for other knowledge mediation channels, such as social proximity and companies’ internal transfers.

One warning note is however due. The harmonization of patent applicants’ names performed by our data sources (EEE-PPAT, as implemented by PatStat) works much better when the applicants are companies from the same country, rather than from different countries. For the latter, it may be often the case that two branches of the same company in different countries are not recognized as such, but treated as different applicants. We have not yet performed a manual check to find out how many of these cases we have in our international sample, and how they may affect our results. If they were many, correcting for them could reduce the estimated marginal effect of social distance, due to the fact that within-company citations are most likely to occur between patents by socially close inventors.

**Figure 2 – Ethnic inventors’ patents: estimated probability of international citation, by co-location, co-ethnicity, and social distance (from specification in table 5c) – India**

![Figure 2](image-url)
5. Discussion and conclusions

By means of patent and inventor data, we have investigated whether social ties binding migrants from the same country help diffusing technical knowledge within the migrants’ destination countries (“diaspora effect”) as well as towards their country of origin (“brain gain effect”). We found evidence of both effects, subject to a number of limitations (both substantive and methodological).

Our empirical exercise has made use of a large and entirely novel sample of patents filed by ethnic inventors in the US, from 1980 to 2010. Ethnic inventors are defined as inventors whose Country of Origin (CoO) falls in a list of 10 countries that the OECD rank among the most important sources of highly skilled migration towards the US. Inventor are assigned to one or another CoO based on linguistic analysis of names and surnames. This technique has a great potential, in that it allows to identify migrant inventors even in the absence of sensitive information such as the inventors’ country of birth or nationality; and it can be extended to other sources of micro-data on highly skilled migrants, such as bibliographic databases (which include scientific authors’ names and surnames).

At the same time, it suffers of major limitation, as it count as “ethnic” both first- and second-generation migrants, as well as descendants of earlier migrants; and that proportions of these categories varies across the CoO considered. Several data quality indicators suggest that our methodology works better for the Asian countries in our list (China, India, Korea, and Japan; with the exception of Iran) than for the European ones (France, Germany, Italy, for Poland; with the possible exception, for the better, of Russia). As such, our methodology suffers of the same limitation as Kerr’s (2007, 2008), although it has the advantage of detailing European CoO, instead of lumping all of them together. We also share with Kerr the difficulties to identify ethnic inventors from English- and Spanish-speaking countries.

We tested for the diaspora effect by sampling all citations by US-resident inventors to the ethnic inventors’ patents in our database (excluding company self-citation and personal self-citations by ethnic inventors), and a set of technology- and year-based controls (JTH methodology). By means of logit regressions, we have tested whether co-ethnicity between inventors increases the probability to observe a citation to occur between patents (diaspora effect), after controlling for the inventors’ location in space (co-location at the BEA level) and onto the network of inventors (social distance).

Our evidence suggests the diaspora effect to be positive and significant for all CoO in our sample, except France, Italy, and Poland. However, this result is not robust to all model specifications. In particular, we show that for all CoO considered, the marginal effect of co-ethnicity is secondary to that of social distance; and that, when we interact the two, it remains positive and significant only for China and India, especially for inventors at more than three degrees of separation. This result goes in the direction of suggesting that the Indian and Chinese diasporas are two special cases, one that the literature suggests to have distinctive properties of size (very large) and cohesion (very tight), which may be expected to be conducive to knowledge diffusion. At the same time, our evidence is not conclusive, for two reasons. First, the weaker results we get for European CoO may depend on the data quality issues discussed above. For this reason, in future version of this paper, we will experiment with different calibrations of the Ethnic-INV algorithm. Second, the sensitiveness of our results for co-ethnicity to controlling for social distance suggests that the former may explain the latter, and this can be true for all CoO. Our further research will investigate to what extent the formation of co-inventorship ties may be positively influenced by co-ethnicity, as suggested by theoretical and empirical work on homophily in network formation (Currarini et al., 2009 and 2010).

In line with Agrawal et al.’s (2008) findings, our estimates for both Indian, but also Chinese inventors suggest that the diaspora effect is a substitute for inventor’s co-location, the marginal effect of ethnic ties on the probability of citation being larger for inventors who are not co-located. We also confirm such marginal effect to be weaker than that of co-location.
As for the brain gain effect, we tested for its presence by sampling all citations by non US-resident inventors to the ethnic inventors’ patents in our database, again adapting the JTH methodology. In this case, we did not exclude company self-citation and personal self-citations by ethnic inventors (returnee inventors). By means of logit regressions, we tested whether inventors who reside the same CoO of US-resident inventors have a higher probability to cite the latter’s patents (Same country dummy variable =1), after controlling for the inventors’ company affiliation (whether they belong to the same company, with branches both in the US and in the focal CoO) and social distance. Evidence is mixed, being it positive and significant for all Asian countries (but Iran) and Russia, but negative or null for the other European countries. When replacing Same country with Co-ethnicity, the estimated coefficients generally increase and turn out to be positive and significant for several European countries. The different results obtained with the two specifications may be explained by the same data quality issues discussed above, but also by the possibility that Co-ethnicity captures some “international diaspora” effects, that is the importance of ethnic ties binding migrant inventors across different destination countries. Future versions of this paper will address the data quality issue, and provide a more specific test for the “international diaspora” effect. Further data issues to be addressed concern the mixed harmonization of patent applicants’ names, which at present may hide many false negatives and lead to under-estimation of company self-citations. Overall, our results are coherent with Kerr’s (2008), with some hints that they could hold for other countries besides China and India.

Besides investigating the role of ethnic ties in the formation of network of inventors, our plans for future research include extending the analysis to Europe as a destination region for intra-regional and extra-regional inventors’ migration. This extension will contribute, among other things, to casting light on a policy–sensitive topic such as the comparative attractiveness of Europe and the US as destinations for migrant scientists and engineers (Cerna and Chou, 2014; Guild, 2007).
Appendix 1 – Inventor names’ disambiguation

A key element of name disambiguation algorithms consists in measuring the edit or phonetic distance between similar names/surnames, and setting some thresholds under which different names/surnames are considered the same (“matching”). Further information contained in the patent documents, as well as benchmarking is then used to validate the matches (“filtering”). Ideally, a good algorithm would minimize both “false negatives” (maximise “recall”) and “false positive” (maximise “precision”). False negatives occur whenever two inventors, whose names or surnames have been spelled or abbreviated differently on different patents, are treated as different persons. False positives occur when homonyms and quasi-homonyms are treated as the same person. Unfortunately, a trade-off exists between the two objectives, which requires making choices based on the consequences of each type of error for the subsequent analysis.

This has two consequences for the analysis of ethnic citations, based upon the linguistic analysis of inventors’ names/surnames:

1) High precision/Low recall algorithms lead to underestimating the number of personal self-citations and overestimating that of co-ethnic citations. This is because all variants of the same inventor’s name and surname will be, most likely, classified as belonging to the same ethnic group. For example, “Vafaie Mehrnaz” and “Vafaie Mehranz” will be both classified as Iranian, but a low recall algorithms may end up treating them as different persons, when instead they are one. The same for the Russian inventor “Yavid Dimitriy”, also appearing on some patents as “Yavid Dimitriy”. When applying the JTH methodology, this problem is magnified by the presence of very prolific inventors, who are responsible for a large number of both cited and citing patents., and thus have the potential to generate a large number of false co-ethnic citations. 24 25

2) When applied to inventor sets from different countries of origin, the same matching rules return different results in terms of pre-filtering precision and recall, due to cross-country differences in the average length of text strings containing names and surnames, and in the relative frequency of common names and surnames 26.

Three complementary strategies may help tackling these problems. The first one consists in making the best possible use of the contextual information contained in patents (that is, to correct for matching errors at the filtering stage). The second consists in using different algorithms to produce more than one datasets, each of which with different combinations of precision and recall, and using them to test the robustness of results. The third one consists in calibrating the disambiguation algorithm by collecting information on linguistic specificities of each country of origin, and exploit them at the matching stage. The information retrieval and computational costs increase when moving from the first to the third strategy. For this reason, our disambiguation algorithm (Massacrator 2.0) does not follow the third one.

---

24 As a further indication of the seriousness the problem, several studies suggest that migrant inventors, at least in the US, are over-represented among prolific ones, due to skill-biased immigration (Stephan and Levin, 2001; Wadhwa et al., 2007a-c; No and Walsh, 2010).

25 High precision/Low recall algorithms may also lead to underestimating the number of returnee inventors. If our Russian inventor patent as “Yavid Dimitriy” and as “Yavid Dimitriy” in Russia, he will not be counted as as a returnee (but his self-citations will be counted as a knowledge flow mediated by ethnicity). However, we suspect this to be a relatively minor problem, as figures of returnee inventors appear too low for their order of magnitude to change with a change in algorithms.

26 Chinese and Korean names and surnames, for example, are both short (which makes it arduous to tell them apart on the sole basis of edit distances) and heavily concentrated on a few, very common ones (such as Wang or Kim). The opposite holds for Russian surnames.
Appendix 2 – Ethnic classification of inventors

When fed with either a name or a surname or both, the IBM-GNR system returns a list of CoAs and two scores of interest:
- “frequency”, which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA;
- “significance”, which approximates the frequency distribution of the name or surname across all CoA.

The IBM-GNR list of CoAs associated to each inventor is too long for being immediately reduced to a unique country of origin for each inventor in our database. This operation requires filtering a large amount of information through an *ad hoc* algorithm, one that compares the frequency and significance of the two lists of CoAs associated, respectively, to the inventor’s name and surname to the inventor’s “country of residence” at the moment of the patent filing (which we obtain from the inventor’s address in the EP-INV dataset). Figure A2.1 illustrates the type of information provided by IBM-GNR, the position of our algorithm in the information processing flow, and the final outcome. Notice that we refer to “country of association” (CoA) when considering the raw information from IBM-GNR, and to “country of origin” when considering the final association between the inventor and one of the many CoAs proposed by IBM-GNR (or one of our “meta-countries” based on linguistic association). The full description of the algorithm is as follows:

I. We consider only inventors in the EP-INV database with at least one patent filed as US residents, or who cite at least one patent filed by US residents, and we assign them to either one of the 10 CoO of our interest, or leave her “unassigned” (which means she may be either a US “native” – whatever it means - or a migrant from other countries)

II. The 10 CoO of our interest are China, India, Iran, Japan, and South Korea (for Asia) and France, Germany, Italy, Poland, and Russia (for Europe). They share two characteristics: they belong to the top 20 CoO of highly skilled migrants in the US, according to OECD/DIOC stock figures for 2005/06 (Widmaier and Dumont, 2011); and their official language is neither English nor Spanish or Portuguese, which is a prerequisite for our algorithm to make sense when applied to migration into the US.

III. For each inventor, we calculate three indicators of her likely CoO:

1. The frequency of her name(s) in English- and Spanish-speaking CoA
2. The product of the significances attached to her name and to the surname, for each CoA coinciding with one of the 10 CoO of our interest. Notice that, in principle, we could find that an inventor is associated to more than one of the 10 countries of interest, either via her name or her surname (for example, a French inventor of Italian descent may have a French name and an Italian surname).

However, these cases are not frequent.

---

27 For example, an extremely common Vietnamese surname such as Nguyen will be associated both to Vietnam and to France, which hosts a significant Vietnamese minority; but in Vietnam it will get a frequency value of 90, while in France it will get only, say, 50, the Vietnamese being just a small percentage of the population. When it comes to significance, the highest percentage of inventors names Nguyen will be found in Vietnam (say 80), followed by France (which has been historically the most important destination countries of Vietnamese migrants besides the US), and several Asian countries, with much smaller values.

28 Language is an issue to the extent that our tools cannot distinguish English-speaking migrant inventors from US ones, nor Spanish-speaking migrants from one country of origin or another. This is why we cannot include in our analysis important origin countries such as the UK, Canada, Mexico and Cuba. We also have not yet included Ukraine and Taiwan, as this will require merging them with Russia and China, respectively. Two other countries in the top 20 list we have not included, for technical reasons, are Vietnam (too few observations among inventors) and Egypt (whose migrants into the US we cannot tell apart from those from other Arab-speaking countries).

29 The intuition is as follows. An inventor with a typical Indian surname, such as Laroia, but named John or Luis is unlikely to be a recent Indian migrant into the US; this is because John and Luis are high-frequency names, respectively, in English-speaking and Spanish-speaking countries. More likely, he will be born in the US from mixed parents; or he could be an Hispanic US citizen, whose surname La Rioja, has been misspelled. On the contrary, Rajiv Laroia is more likely to be a first- or second-generation Indian immigrant, as Rajiv is high-frequency name in India, a zero-frequency name in Spanish-speaking countries, and a low-frequency name in English-speaking countries (some of which host sizeable communities of Indian descent).
3. The significance attached to the surname in the CoA associated to indicator n.2. 30

As a result, we will have, for each inventor, one (or very few) candidate CoO and three indicators of potential success of this “candidacy”.

IV. We set six possible threshold values for indicator n.1 (from 10 to 100, with steps of 20), eleven threshold values for indicator n.2 (from 0 to 10000, with steps of 1000), and six threshold values for indicator n.3 (from 50 to 100, with steps of 10). We consider 102 combinations of such threshold values (“calibrations”), and for each combination we assign each inventor to one or another CoO (or to no CoO at all). Each inventor is therefore associated to one vector of 102 dummies (one for each calibration) and a specific CoO, with dummy=1 indicating that the inventor comes for that CoO, and dummy=0 that she does not (no CoO assigned). 31

V. We apply steps I. to IV. also to inventors in the WIPO-PCT database by Fink and Miguelez (2013), which report the inventors’ nationality, which we use as benchmark to evaluate the precision and recall rates obtained by each calibration, for each CoO. We then identify Pareto-optimal calibration, namely the calibrations whose precision rate cannot be improved upon without losing out on the recall rate, and viceversa (blue dots in figures A2.2, which report the calibration results for China and Italy). Notice that the Pareto-optimal calibrations are not necessarily the same for all CoO; again from figure A2.2, one can the Pareto-Optimal calibrations for China are more convex than those for Italy. In other words, they imply a much less sharp trade-off between precision and recall: while for Italy we can attain a 70% precision rate only at the cost of reducing the recall rate to 10%, for China we reduce the latter only to 60%. The precision-recall trade-off can be considered a measure of the quality of our algorithm, per country. In general, such quality is higher for Asian countries (with the exception of Iran) than for European ones.

VI. Finally, we retain for our analysis two calibrations per CoO: a “high recall” calibration (one that ensures the highest recall value, conditional on precision being at least 30%); and a “high precision” calibration, one that requires precision to be no less than 70%. High recall values are expected to include a large number of false positives (inventors wrongly assigned to one or another of the 10 CoO of interest), but also to accommodate for a looser definition of migrant inventors, one that includes many second- and third-generation migrants. The latter’s validity depends on the strength of ties binding such migrants to other US residents of the same descent, and/or to their countries of origin (on which we have no a priori information).

In the present version of the paper, we make use only of “high recall” calibration results. To further compare data quality across CoO, we can inspect the frequency distribution of values taken by indicator n.2 (figure A2.3). The more right (left) skewed the distribution, the better (worse) the quality: the most striking comparison here is between India and Italy, with the former clearly exhibiting higher quality. According to this measure, too, quality is generally higher for Asian countries (with the exception of Iran) than for European ones.

---

30 The intuition is as follows: the indicator n.2 may have a high value due exclusively to a very high value of the significance for the name, with a moderate value for the significance of the surname. We wish the latter not to be too low.

31 Keeping with the example from the previous footnotes, Rajiv Laroia will be associated to CoO=India, with a vector containing n<102 zeroes and 102-n ones. The ones are all associated with “high recall” combinations of high threshold values for indicator n.1 and low threshold values for nr.2 and nr.3 (such as, respectively, 70-5000-60; see figure 1), while the zeroes will be associated with “high precision” combinations (low threshold values for indicator n.1 and high threshold values for nr.2 and nr.3; such as, respectively, 30-8000-80). Rajiv Laroia will be confirmed having CoO=India only in the high recall case, but not in the high precision case (for which indicator nr.1 is too high). In practice, the high precision combination leaves the door open to Rajiv Laroia’s CoO being the UK, and to Rajiv Laroia being possibly of Indian descent, but with no ties to India or to Indian migrants in the US.
Figure A2.1 From inventor data to the Ethnic-INV database

1) General workflow

2) Details of Ethnic-INV algorithm

<table>
<thead>
<tr>
<th>Surname</th>
<th>Country of Association</th>
<th>Frequency</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAROIA</td>
<td>INDIA</td>
<td>10</td>
<td>99</td>
</tr>
<tr>
<td>LAROIA</td>
<td>FRANCE</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First name</th>
<th>Country of Association</th>
<th>Frequency</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAJIV</td>
<td>INDIA</td>
<td>50</td>
<td>01</td>
</tr>
<tr>
<td>RAJIV</td>
<td>GREAT BRITAIN</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>RAJIV</td>
<td>SRI LANKA</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>RAJIV</td>
<td>TRINIDAD</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>RAJIV</td>
<td>AUSTRALIA</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>RAJIV</td>
<td>CANADA</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>RAJIV</td>
<td>NETHERLANDS</td>
<td>50</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country of Association</th>
<th>Max frequency of first name in Anglo/Hispanic countries (1)</th>
<th>JOINT Significance (2)</th>
<th>Significance of surname (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INDIA</td>
<td>50</td>
<td>8019</td>
<td>99</td>
</tr>
<tr>
<td>FRANCE</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GREAT BRITAIN</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SRI LANKA</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TRINIDAD</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AUSTRALIA</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CANADA</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NETHERLANDS</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Thresholds (India-specific)

<table>
<thead>
<tr>
<th>High Recall</th>
<th>High Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>5000</td>
</tr>
<tr>
<td>30</td>
<td>8000</td>
</tr>
</tbody>
</table>

Do indicators (1)-(5) pass all thresholds?
- Yes
- No
Figure A2.2a - Ethnic-INV algorithm calibration results: China and Italy
Figure A2.3 - Frequency distribution of values taken by indicator n.2: India vs. Italy

Distribution of joint significance - Indians

Distribution of joint significance - Italians
Appendix 3 – Additional tables

Figure A3.1 – Share of ethnic inventors of EPO patent applications by US residents; by CoO

Figure A3.2 – Ethnic inventors’ share of EPO patent applications by US residents; by CoO
Figure A3.5

Location Quotient - French

Figure A3.6

Location Quotient - Indians
References


Saxenian, A., Motoyama, Y., Quan, X., 2002, Local and global networks of immigrant professionals in Silicon Valley. Public Policy Instit. of CA.


